# 6. DISTRIBUCIONES MUESTRALES

## CONTENIDO

# 6 DISTRIBUCIONES MUESTRALES

## 6.1 INTRODUCCIÓN

En el capítulo 1 hemos definido la **inferencia estadística** como un proceso que usa información proveniente de la muestra para generalizar y tomar decisiones acerca de toda la población en estudio. Sin embargo, hasta el momento hemos trabajado la muestra y la población por separado.

En el capítulo 2, trabajamos herramientas útiles en el análisis exploratorio de los datos provenientes de una muestra, tanto gráficos como resúmenes numéricos para extraer información de interés para la inferencia. Hablamos de distribuciones de frecuencias y estadísticos.

En los capítulos 3, 4 y 5, a través del lenguaje de la probabilidad, tratamos los modelos para las poblaciones que pueden ser de interés, sobre las cuales nos interesa sacar conclusiones, o tomar una decisión. Definimos las variables aleatorias, sus distribuciones de probabilidad, parámetros y algunos modelos frecuentes.

Podemos hacer un cuadro comparativo entre características del análisis exploratorio de datos y de la inferencia estadística:

| *Análisis exploratorio de datos* | *Inferencia estadística* |
|---|---|
| Su objetivo es la exploración de los datos muestrales, en busca de regularidades interesantes. | Su objetivo es responder preguntas concretas sobre la población, planteadas antes de la obtención de los datos. |
| Las conclusiones sólo se aplican a las unidades de análisis y a las circunstancias para las cuales se obtuvieron los datos. | Las conclusiones se extienden a toda la población en estudio. |
| Las conclusiones se basan en lo que “vemos” en los datos. | Las conclusiones se explicitan con un grado de confianza. |

Muchas de las técnicas utilizadas en inferencia exigen, también, que la distribución de los datos tenga determinadas características. El análisis de datos es de gran ayuda en este aspecto, para descubrir observaciones atípicas y otras desviaciones que puedan perturbar una correcta inferencia. Por lo tanto, en la práctica podemos observar cómo el análisis exploratorio de los datos y la inferencia estadística se complementan.

Como se mencionó en el capítulo I, muy frecuentemente es necesario seleccionar una **muestra** de unidades de la población, para extraer conclusiones respecto de la misma, en base a las observaciones muestrales (ver Muestra, pag. 7).

Sintetizando:

Cuando el interés reside en generalizar las conclusiones de los resultados observados a la población en estudio o queremos tomar una decisión sobre la población en base a una muestra, estamos frente a un problema de inferencia estadística.

Para que este proceso sea adecuado, debemos tener en cuenta:

- Plantear claramente el problema.
- Delimitar la población en estudio[1]
- Definir si el objetivo reside en estimar el valor de un parámetro desconocido de la población (por ej. $\mu$, $\sigma$, p) a partir de un ***estadístico*** calculado con los datos de una muestra o decidir sobre valores hipotéticos que asignamos a dichos parámetros.
- Hacer un correcto diseño para la obtención de los datos muestrales [2]. Los resultados de las técnicas para la inferencia que se utilizarán sólo serán válidos si la muestra es obtenida por métodos aleatorios, que son los métodos que dan "confianza" de seleccionar **muestras representativas** de la población. Un buen diseño para la obtención de los datos, es la mejor garantía de que la inferencia tenga valor.
- Tener en cuenta y verificar los requerimientos de las técnicas a aplicar

## 6.2 PARÁMETROS Y ESTADÍSTICOS

Un **parámetro** es un número que describe algún aspecto de la población en estudio. En la práctica, en la mayoría de los casos (población infinita, pruebas destructivas, etc) el valor del parámetro es desconocido.

Un **estadístico** es un número que se calcula a partir de los datos muestrales. Si se lo utiliza para estimar un parámetro desconocido, se lo conoce con el nombre de **estimador**.

El objetivo de este capítulo y los próximos es desarrollar el **por qué y cómo** se utilizan los estadísticos para estimar a los correspondientes parámetros.

Tengamos en cuenta que el valor del parámetro es fijo, mientras que el valor de un estadístico está en función de la muestra seleccionada y por lo tanto podrá variar de una muestra a otra.

Si de alguna manera, pudiéramos medir la precisión de este proceso, es decir, si pudiéramos evaluar si el valor del estadístico va a estar cerca del valor del parámetro correspondiente, para cualquier muestra extraída de la población, entonces estaríamos en condiciones de hacer "buenas inferencias". Es aquí donde la técnica de muestreo y el tamaño de la muestra juegan un papel fundamental.

Como se mencionó en el capítulo I, pag. 8 trabajaremos con muestras aleatoria simples, en donde cada elemento de una muestra de tamaño n es una variable aleatoria, siendo $X_1$, $X_2$,… $X_n$, variables independientes entre sí[3].

Sólo cuando se utiliza el azar para escoger los elementos que conforman una muestra, podemos describir cómo varía el estadístico. Al obtener de forma repetida de una población distintas muestras del mismo tamaño, podemos encontrar la ***distribución muestral del estadístico***, como veremos seguidamente.

---

[1] Así, por ej., si se quiere analizar cierta característica de los alumnos que cursaron en la Fac. Reg. Rosario de la UTN en los años 2005 y 2006 y la muestra se elige seleccionando alumnos al azar solamente de los que cursaron durante esos años en ISI, las conclusiones que se extraigan a partir de esta muestra serán válidas **sólo** para la población de los alumnos de ISI, pero no para **todos** los alumnos de la Fac. Reg. Rosario.

[2] Ya hemos dicho que en el curso sólo se trabajará con **muestras aleatorias simples**.

[3] En el caso de poblaciones finitas, el muestreo debe ser con reposición para que la ocurrencia de una observación no aumente o disminuya la probabilidad de ocurrencia de otra (ver ej. 1 y 2, pag. 58, Cap. III)

Denominamos variabilidad muestral al hecho de que el valor de un estadístico varía en un muestreo aleatorio repetido. Para comprenderlo podemos recurrir a la teoría de la probabilidad o la simulación a partir de un ejemplo sencillo, como el que se plantea a continuación.

## 6.3 DISTRIBUCION DEL PROMEDIO MUESTRAL

Consideramos:

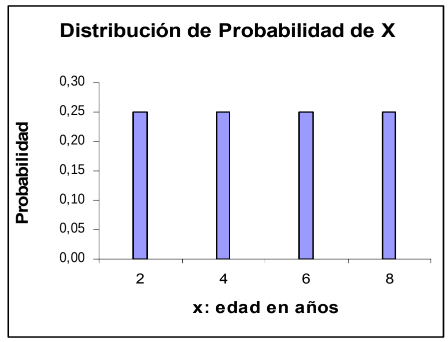
- Variable aleatoria → X: edad de hermanos de una familia, en años

  Tamaño de la población → N = 4

  Espacio muestral → $S_x = \{2, 4, 6, 8\}$

  Distribución de probabilidad → Uniforme discreta

| Tabla 1 | |
|---|---|
| x | P(X=x) |
| 2 | 0,25 |
| 4 | 0,25 |
| 6 | 0,25 |
| 8 | 0,25 |
| | 1 |




Parámetros: Esperanza Matemática $\mu_x = 5$
Varianza $\sigma^2_x = 5$

- Tomamos **muestras de tamaño 2, con reposición**. Cada muestra es de la forma ($x_1$ ; $x_2$), donde:

  $X_i$ es el i-ésimo elemento de la muestra.

  Simbolizaremos con $\overline{X}$ al promedio muestral y con *S* al desvío estándar de la muestra.

  En la tabla 2 están todas las posibles muestras con sus correspondientes promedios.

  Observamos que tanto el primer elemento de la muestra como el segundo son variables aleatorias, ya que, antes de realizar el muestreo, no sabemos qué valores tomarán. Si consideramos la distribución de probabilidad de cada una de ellas, resultan idénticas a la distribución de probabilidad de la población de la cual fueron extraídas las muestras, siendo por lo tanto, iguales los parámetros estadísticos:

  $$\mu_{xi} = 5 \quad \text{y} \quad \sigma^2_{xi} = 5$$

  En la tabla 2, también se visualiza que el promedio muestral es una variable aleatoria.

| Tabla 2 | | |
|---|---|---|
| Muestra, n=2 | | |
| $x_1$ | $x_2$ | Promedio |
| 2 | 2 | 2 |
| 2 | 4 | 3 |
| 2 | 6 | 4 |
| 2 | 8 | 5 |
| 4 | 2 | 3 |
| 4 | 4 | 4 |
| 4 | 6 | 5 |
| 4 | 8 | 6 |
| 6 | 2 | 4 |
| 6 | 4 | 5 |
| 6 | 6 | 6 |
| 6 | 8 | 7 |
| 8 | 2 | 5 |
| 8 | 4 | 6 |
| 8 | 6 | 7 |
| 8 | 8 | 8 |

Consideramos ahora esta nueva variable aleatoria

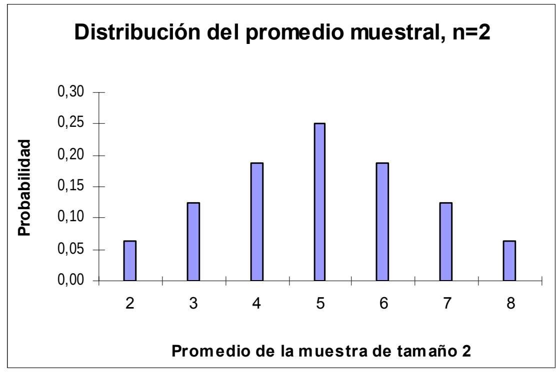
$\overline{X}$ : edad promedio de 2 hermanos elegidos al azar de entre los 4, con reposición.

Como podemos observar en la Tabla 2, para la variable $\overline{X}$ :

Espacio muestral $S_{\overline{x}}$ = {2, 3, 4, 5, 6, 7, 8}

La distribución de probabilidad es:

| Tabla 3: n=2 | |
|---|---|
| $\overline{X}$ | *Probabilidad* |
| 2 | 0,0625 |
| 3 | 0,1250 |
| 4 | 0,1875 |
| 5 | 0,2500 |
| 6 | 0,1875 |
| 7 | 0,1250 |
| 8 | 0,0625 |
| | 1,0000 |




Parámetros: Esperanza Matemática $\mu_{\overline{x}}$ = 5

Varianza $\sigma^2_{\overline{x}}$ = 2.5

Observemos que la distribución del promedio adopta una forma completamente distinta de la distribución uniforme de los datos de origen.

- Repitiendo la experiencia, con muestras de tamaños 3, 4 y 5 respectivamente, obtenemos las distribuciones de los promedios que mostramos a continuación, acompañadas de las gráficas respectivas:

n = 3

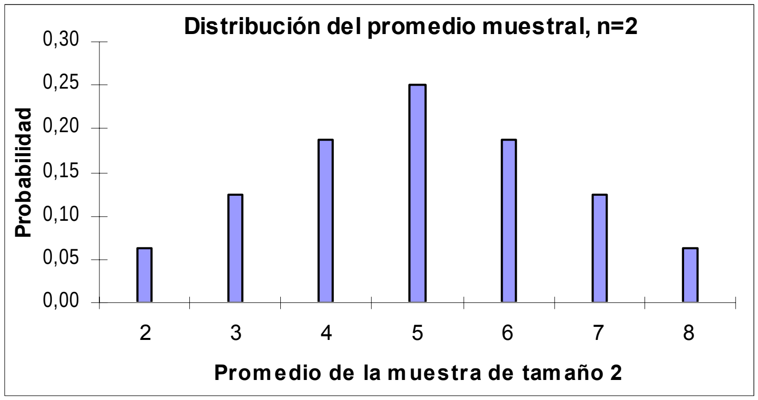
| *Promedio* | *Probabilidad* |
|---|---|
| 2,00 | 0,015625 |
| 2,67 | 0,046875 |
| 3,33 | 0,093750 |
| 4,00 | 0,156250 |
| 4,67 | 0,187500 |
| 5,33 | 0,187500 |
| 6,00 | 0,156250 |
| 6,67 | 0,093750 |
| 7,33 | 0,046875 |
| 8,00 | 0,015625 |
| | 1 |

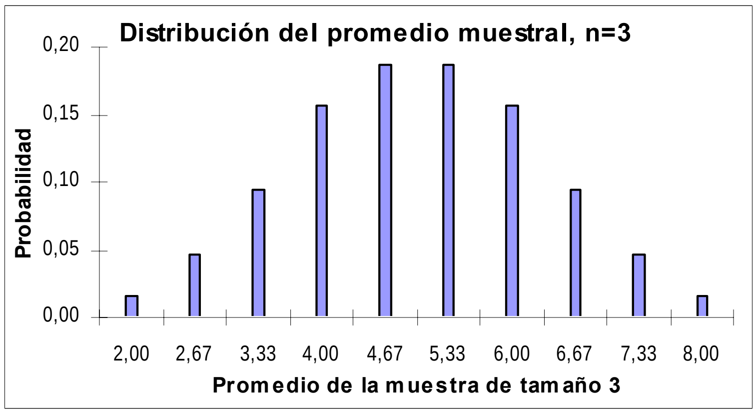
n = 4

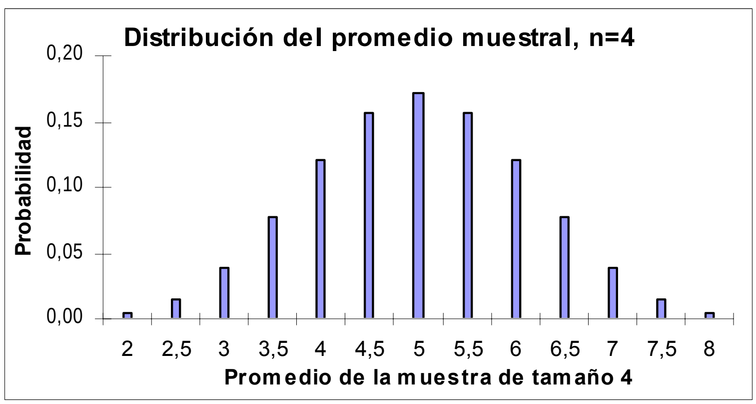
| *Promedio* | *Probabilidad* |
|---|---|
| 2 | 0,00390625 |
| 2,5 | 0,01562500 |
| 3 | 0,03906250 |
| 3,5 | 0,07812500 |
| 4 | 0,12109375 |
| 4,5 | 0,15625000 |
| 5 | 0,17187500 |
| 5,5 | 0,15625000 |
| 6 | 0,12109375 |
| 6,5 | 0,07812500 |
| 7 | 0,03906250 |
| 7,5 | 0,01562500 |
| 8 | 0,00390625 |
| | 1 |

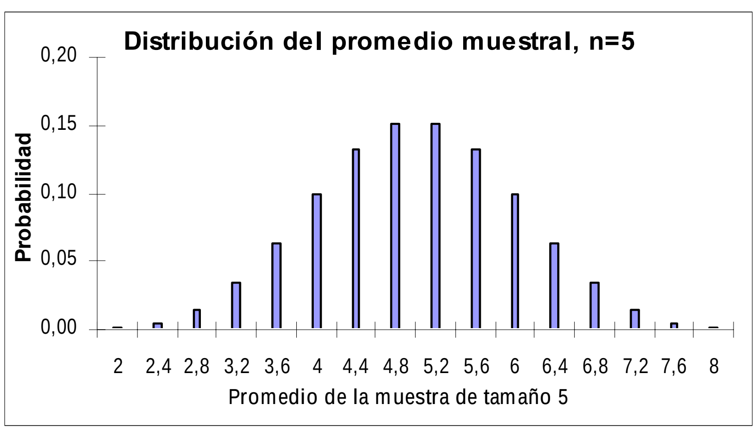
n = 5

| *Promedio* | *Probabilidad* |
|---|---|
| 2 | 0,00097656 |
| 2,4 | 0,00488281 |
| 2,8 | 0,01464844 |
| 3,2 | 0,03417969 |
| 3,6 | 0,06347656 |
| 4 | 0,09863281 |
| 4,4 | 0,13183594 |
| 4,8 | 0,15136719 |
| 5,2 | 0,15136719 |
| 5,6 | 0,13183594 |
| 6 | 0,09863281 |
| 6,4 | 0,06347656 |
| 6,8 | 0,03417969 |
| 7,2 | 0,01464844 |
| 7,6 | 0,00488281 |
| 8 | 0,00097656 |
| | 1 |


Distribución del promedio muestral, n=2
Probabilidad
0,30
0,25
0,20
0,15
0,10
0,05
0,00
2
3
4
5
6
7
8
Promedio de la muestra de tamaño 2


Distribución del promedio muestral, n=3
Probabilidad
0,20
0,15
0,10
0,05
0,00
2,00
2,67
3,33
4,00
4,67
5,33
6,00
6,67
7,33
8,00
Promedio de la muestra de tamaño 3


Distribución del promedio muestral, n=4
Probabilidad
0,20
0,15
0,10
0,05
0,00
2
2,5
3
3,5
4
4,5
5
5,5
6
6,5
7
7,5
8
Promedio de la muestra de tamaño 4


Distribución del promedio muestral, n=5
Probabilidad
0,20
0,15
0,10
0,05
0,00
2
2,4
2,8
3,2
3,6
4
4,4
4,8
5,2
5,6
6
6,4
6,8
7,2
7,6
8
Promedio de la muestra de tamaño 5

En las gráficas anteriores podemos comprobar una aplicación del *Teorema Central del Límite*: a medida que aumenta el tamaño de la muestra la distribución de probabilidad del promedio muestral se hace cada vez más acampanada, concentrándose alrededor del promedio de la población original.

La Tabla 4 permite comparar los parámetros esperanza matemática y varianza de la población original, con la esperanza matemática y varianza de las poblaciones de los promedios muestrales antes descriptas.

Vemos que las medias poblacionales se mantienen iguales a 5 (la esperanza matemática de las edades de los 4 hermanos), mientras que las varianzas poblacionales disminuyen su valor a medida que aumenta el tamaño de la muestra.

Tabla 4

| Población | $\mu_x = 5$ | $\sigma_x^2 = 5$ |
|---|---|---|
| Tamaño de muestra | $\mu_{\overline{x}}$ | $\sigma_{\overline{x}}^2$ |
| 2 | 5 | 2,5 |
| 3 | 5 | 1,667 |
| 4 | 5 | 1,25 |
| 5 | 5 | 1 |

Al considerar la distribución de los valores tomados por el estadístico $\overline{\mathbf{X}}$ en todas las muestras de un mismo tamaño **n** de la misma población, obtenemos la **distribución muestral de $\overline{\mathbf{X}}$.**

Generalizando:

***Distribución muestral de la media muestral $\overline{X}$***

Si las muestras aleatorias simples de tamaño n son tomadas de una población con media poblacional $\mu$ y desvío estándar poblacional $\sigma$, la distribución muestral de $\overline{X}$ tiene las siguientes propiedades:

⇨ 1) $\mu_{\overline{X}} = E(\overline{X}) = \mu$

Es decir, el promedio de todos los posibles valores de $\overline{X}$ es igual al parámetro $\mu$

⇨ 2) $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

Cuando el tamaño de la muestra aumenta, la medida de dispersión disminuye. Es decir, a medida que el número de observaciones obtenidas aumenta, el promedio de los valores observados se acerca más y más a $\mu$ (Ley de los grandes números)

⇨ 3) Si la población de la cual se extraen las muestras es normal, la distribución de $\overline{X}$ es también normal con media y desvío como los dados en los puntos anteriores, para cualquier tamaño muestral n.

⇨ 4) Si la población de la cual se extraen las muestras no es normal, pero el tamaño muestral es "suficientemente" grande, la distribución de $\overline{X}$ es aproximadamente normal con media y desvío como los dados en los puntos anteriores. Suficientemente grande en la práctica significa un tamaño de muestra n ≥ 30 (Teorema Central del Límite).

El tamaño n de la muestra, necesario para que $\overline{X}$ se aproxime a una distribución normal depende de la distribución de la población. En el caso de que las muestras se extraigan de una población uniforme son suficiente 6 observaciones para que la distribución del promedio muestral sea aproximadamente normal.

⇨ 5) Si la población de la cual se extraen las muestras es normal, con media poblacional $\mu$ y desvío estándar poblacional $\sigma$, pero ésta es desconocida, se reemplaza $\sigma$ por **S** (desvío estándar muestral) y la estadística $\dfrac{(\overline{x}-\mu)}{S/\sqrt{n}}$ deja de tener distribución normal estandarizada y tiene una distribución *t Student con n-1 grados de libertad* [(a)] :

$$\frac{(\overline{X}-\mu)}{S/\sqrt{n}} \sim t_{n-1;\alpha}$$

*(Ver demostraciones en el Apéndice)*

[(a)] La apariencia general de la distribución t es similar a la de la distribución normal estándar: ambas son simétricas y unimodales y el valor máximo de la ordenada se alcanza en la media $\mu$ = 0. Sin embargo esta distribución tiene colas más amplias que la normal. Existe una distribución *t* distinta para cada tamaño de muestra. Una distribución *t* viene determinada por un parámetro llamado grados de libertad. A medida que aumentan los grados de libertad, la curva de densidad *t* se parece más a la curva de la N(0,1), ya que la estimación de $\sigma$ por s se va haciendo más precisa.

- La propiedad 1 indica que el estimador $\overline{\mathbf{X}}$ es ***insesgado***, ya que el centro de su distribución muestral es igual al valor del parámetro poblacional correspondiente.

- La propiedad 2 hace a la **variabilidad o precisión** del estimador y vemos que a medida que el tamaño muestral crece la precisión del estimador es mayor, ya que la variación alrededor del parámetro desconocido disminuye (propiedad de ***convergencia***). Si la distribución de un estadístico muestra valores muy alejados, se dice que carece de precisión.

*Idealmente buscamos un estimador que cumpla estas dos propiedades: que sea insesgado y convergente[4]:*

> Un estadístico es **insesgado** si el centro de su distribución muestral es igual al valor del parámetro poblacional correspondiente.
>
> Un estadístico es **convergente** si su desviación estándar disminuye a medida que el tamaño de muestra crece.

El estadístico $\overline{\mathbf{X}}$, por poseer estas propiedades, es un buen estimador de $\boldsymbol{\mu}$.

Estas propiedades también se cumplen para la proporción muestral o frecuencia relativa ( $\mathbf{f_r}$ ) y la varianza muestral ($\mathbf{S^2_{n-1}}$), siendo por lo tanto respectivamente, buenos estimadores de la proporción poblacional (**p**) y varianza poblacional ($\sigma^2$), como veremos en los puntos 6.4 y 6.5.-

En general, la notación que utilizaremos para los estimadores es la siguiente:

| Parámetro | Estimador |
|---|---|
| $\mu$ | $\hat{\mu} = \overline{\mathbf{X}}$ |
| p | $\hat{p} = f_r$ |
| $\boldsymbol{\sigma}^2$ | $\hat{\sigma}^2 = S^2_{n-1}$ |

4 Estas condiciones permiten controlar los errores de estimación al aumentar el tamaño de la muestra, como veremos más adelante.

## 6.4 DISTRIBUCION DE LA FRECUENCIA RELATIVA o PROPORCIÓN MUESTRAL

El estadístico $\hat{p} = f_r$ es un buen estimador del parámetro **p** (proporción poblacional o probabilidad).

Si simuláramos tomar muchas muestras de igual tamaño y en cada una de ellas calculáramos la proporción de veces que ocurre un suceso A, hallaríamos:

- ⇒ La distribución de la proporción muestral es aproximadamente normal [5]
- ⇒ Su media se encuentra cerca de la proporción poblacional p
- ⇒ Su desviación estándar se hace menor a medida que el tamaño de la muestra se hace mayor.

Generalizando:

***Distribución muestral de la*** $\hat{p} = f_r$ (proporción muestral)

Si de una población donde p representa la proporción de elementos que tienen cierta característica A, se toman muestras aleatorias simples de tamaño n, la distribución muestral de la proporción muestral o frecuencia relativa ($\hat{p} = f_r$) de las veces que ocurre A en n, tiene las siguientes propiedades:

⇨ 1) **E ($f_r$) = p**

Es decir, el promedio de todos los posibles valores de $f_r$ es igual al parámetro **p**.

⇨ 2) $$\sigma_{\hat{p}} = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

Cuando el tamaño de la muestra aumenta, la medida de dispersión disminuye. Es decir, a medida que el número de observaciones obtenidas aumenta, el promedio de los valores observados se acerca más y más a p (Ley de los grandes números).

Observe que para un tamaño de muestra fijo, la máxima desviación estándar se encuentra en p = 0,5

⇨ 3) Si n es "suficientemente" grande [(b)], la distribución de $\hat{p} = f_r$ se comporta aproximadamente como una distribución normal con media y desviación estándar como las dadas en los puntos 1 y 2.

$$\hat{p} \text{ es aproximadamente } N\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$$

*(Ver demostraciones en el Apéndice)*

---

⇒ [5] Para poder aproximar la distribución Binomial a la Normal, el tamaño de muestra n debe ser "suficientemente" grande. Como regla empírica esta aproximación es apropiada si np > 5 (Cap. 4).

## 6.5 DISTRIBUCION DE LA VARIANZA MUESTRAL

El estadístico $\mathbf{S}^2_{\mathbf{n}-1}$ es un buen estimador del parámetro $\sigma^2$ (varianza poblacional).[6]

Si simuláramos tomar muchas muestras de igual tamaño y en cada una de ellas calculáramos la varianza muestral, hallaríamos:

⇒ La media de la varianza muestral se encuentra cerca de la varianza poblacional $\sigma^2$

⇒ Su desviación estándar se hace menor a medida que el tamaño de la muestra se hace mayor.

Generalizando:

***Distribución muestral de la*** $\mathbf{S}^2$ (varianza muestral)

Si de una población se toman muestras aleatorias simples de tamaño n, la distribución muestral de la varianza muestral $\mathbf{S}^2_{\mathbf{n}-1}$, tiene las siguientes propiedades:

⇨ 1) $E(S^2) = \sigma^2$

Es decir, el promedio de todos los posibles valores de $\mathbf{S}^2_{\mathbf{n}-1}$ es igual al parámetro $\sigma^2$

⇨ 2) $V(S^2) = \sigma^2_{S^2} = \dfrac{2\sigma^4}{n-1}$

Cuando el tamaño de la muestra aumenta, la medida de dispersión disminuye. Es decir, a medida que el número de observaciones obtenidas aumenta, el promedio de los valores observados de $S^2$ se acerca más y más a $\sigma^2$ (Ley de los grandes números).

⇨ 3) Si la población de la cual se extraen las muestras es normal, la variable $\dfrac{(n-1)\,S^2}{\sigma^2}$ tiene una distribución ji cuadrado ($\chi^2$) con $n-1$ grados de libertad [(b)]:

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

⇨ 4) Si n es “suficientemente” grande, la distribución de la variable $\chi^2$ se ve como una distribución normal con media y desviación estándar como las dadas en los puntos 1 y 2.

*(Ver demostraciones en el Apéndice)*

[(b)] Las distribuciones ji cuadrado son una familia de distribuciones que sólo toman valores positivos y que son asimétricas hacia la derecha. Una distribución ji cuadrado viene determinada por un parámetro llamado grados de libertad. A medida que aumentan los grados de libertad, las curvas de densidad son menos asimétricas y por lo tanto, los valores mayores son más probables.

[6] Utilizaremos la notación $S^2$ para identificar a la variable $S^2_{n-1}$ (varianza muestral).

En este material hemos tratado el comportamiento de las distribuciones muestrales de algunos estimadores cuando se toman muestras aleatorias simples.

Se analizó que si el tamaño de muestra es más grande, la distribución de estos estimadores tiende a centrarse más y más alrededor del valor del parámetro que se quiere estimar.

En la práctica no se conocerá el verdadero parámetro poblacional (por eso la estimación) y se tomará una sola muestra (no muchas como cuando se simuló la distribución del promedio muestral), pero son las propiedades (insesgado y convergencia) las que garantizan que cuando la muestra que se toma sea grande habrá una alta probabilidad de que el valor que toma el estimador (estimación) esté cerca del verdadero valor del parámetro que se quiere estimar.

## 6.6 EJERCICIOS PROPUESTOS [7]

1.- El 9 % de los individuos de una región tiene sangre tipo B. En una muestra simple al azar de 400 personas de esa población se encontró que 12,5 % tenían sangre tipo B.

a) Indique:

- valor numérico del parámetro: …….
- valor numérico del estadístico: …….
- identifique en términos del problema al parámetro y al estadístico

b) ¿Cuál es la probabilidad de que una nueva muestra aleatoria de tamaño 400 contenga por lo menos un porcentaje de 12,5 % de personas con sangre tipo B?

2.- Considere la variable aleatoria X: peso de alumnos varones de UTN, FRRO.
Se conoce que esta variable tiene una distribución normal con promedio 75 kg y una desviación estándar de 7 kg.

a) Grafique y compare las distribuciones muestrales de $\overline{X}$ cuando se extraen muestras aleatorias simples de:

* 10 alumnos * 30 alumnos * 100 alumnos

b) ¿Cuál es la proporción de muestras de tamaño 30 que arrojarán un valor del promedio alejado del promedio poblacional en a lo sumo 2 desviaciones estándares?

3.- Suponga que el 60% de todos los estudiantes de la UTN, Reg. Rosario acceden a información sobre cursos por medio de Internet.

a) Grafique en forma aproximada la distribución para la posible proporción muestral basada en una muestra aleatoria simple de 100 estudiantes.

b) ¿Cuál es la probabilidad de observar una proporción muestral de 0,50 basada en una muestra aleatoria simple de tamaño 100 si la proporción poblacional fuese de 0,60? Explique.

4.- Sea X el número de accidentes por semana en una esquina dada. Suponga que la media de X es 2,2 y el desvío estándar de X es 1,4.

a) Sea $\overline{X}$ el número promedio de accidentes por semana en un año, o sea, n= 52 semanas. ¿cuál es la distribución aproximada de la media muestral? Bosquéjela.

[7] Los ejercicios 1, 3 y 4 fueron extraídos y adaptados del módulo Número 7: Distribuciones muestrales de la Colección Métodos Estadísticos I., redactado por docentes de la UNR y extractado del libro "Interactive Statistics" de Martha Aliaga, Universidad de Michigan, 2002.

b) ¿cuál es la probabilidad de que el promedio de accidentes por semana en un año sea menor que 2?

c) ¿Cuán probable es que el número total de accidentes por año sea menor que 100?. Sugerencia: P (Total $<$ 100) = P (Promedio $< \frac{100}{52}$).

5.- Un contratista de obras viales ha tomado un contrato para construir una carretera de hormigón de 800 km. de longitud. La carretera a construir será inspeccionada por vialidad tomando muestras de tamaño 9 por cada 5 km. construidos. El tramo de 5 km. se aceptará sin objeciones si la media de los espesores de las 9 determinaciones supera 149 mm. En este caso la ganancia para el contratista es de 5000$ en los 5 km.
Si el promedio de las 9 determinaciones de los espesores se encuentra entre 140 mm y 149 mm también se aprueba el tramo pero con una quita en el precio, obteniéndose entonces una utilidad de 2300$.
En cualquier otro caso se debe rehacer el tramo, lo que significa una pérdida de 3000$.
La variable aleatoria espesor tienen distribución normal con esperanza matemática igual a 145 mm y desvío estándar igual a 15 mm.
Considere la variable aleatoria "utilidad por tramo de 5 km." Encuentre su distribución de probabilidad, esperanza matemática y varianza.

¿Cuál es la probabilidad de que la construcción de los 800 km. dé una utilidad inferior a 300.000$ si los espesores son independientes de tramo a tramo?

6.- Demuestre que $\overline{\mathbf{X}}$ es un estimador consistente de $\mu$ mediante el empleo de la desigualdad de Chebyshev (vea ley de los grandes números en apéndice)

# APÉNDICE

## 1.- Promedio Muestral o Media Aritmética

$$\overline{X} = \frac{1}{n}\sum_{1}^{n} Xi$$

Siendo E(X) = $\mu$ y V(X) = $\sigma^2$

y teniendo una Muestra Aleatoria Simple de tamaño n $\Rightarrow$ ($X_1$, $X_2$,.....$X_n$),

entonces E( $X_i$ ) = $\mu$ y V($X_i$) = $\sigma^2$ (estas condiciones fueron visualizadas en pag. 129 para muestras de tamaño 2).

Luego E ( $\overline{X}$ )= E $[\frac{1}{n}\sum_{1}^{n} Xi] = \frac{1}{n}E\langle\sum_{1}^{n} Xi\rangle = \frac{1}{n}\sum_{1}^{n} E(Xi) = \frac{1}{n}n\mu = \mu$

V ( $\overline{X}$ )=V $[\frac{1}{n}\sum_{1}^{n} Xi] = \frac{1}{n^2}V\langle\sum_{1}^{n} Xi\rangle = \frac{1}{n^2}\langle\sum_{1}^{n} VXi\rangle = \frac{1}{n^2}nV\langle Xi\rangle = \frac{1}{n}\sigma^2$

Con respecto a la distribución la variable **promedio muestral** podemos decir:

a) Si la variable aleatoria X se distribuye normalmente, por propiedad reproductiva :

$$\overline{X} \sim N\langle\mu, \frac{\sigma}{\sqrt{n}}\rangle$$

b) Si la variable X tiene cualquier distribución pero n $\geq$ 30 aplicando el Teorema Central del Límite, la distribución aproximada es :

$$\overline{X} \sim N\langle\mu, \frac{\sigma}{\sqrt{n}}\rangle$$

## 2.- Proporción Muestral o Frecuencia Relativa

Sea una experiencia aleatoria y un suceso A asociado a la misma. Se realizan n repeticiones independientes de la experiencia. Se definen las variables:

$n_A$ : número de veces que ocurre A en las n repeticiones independientes de la experiencia

$f_A$ :proporción de veces que ocurre A en las n repeticiones independientes de la experiencia

siendo $f_A = \frac{n_A}{n}$

La variable aleatoria $n_A$ tiene distribución binomial con E ($n_A$) = n.p y $\boldsymbol{\sigma^2}$ ($n_A$) = n.p.(1-p)

y en consecuencia $\boxed{E(f_A) = p}$ y $\boxed{\sigma^2(f_A) = \frac{p(1-p)}{n}}$

## 3.- **Varianza muestral** ($S^2$)

En la página 135 se plantea, para el caso en que la variable X se distribuya normalmente, que:

$$E(S^2) = \sigma^2$$

$$V(S^2) = \sigma^2_{S^2} = \frac{2\sigma^4}{n-1}$$

A continuación se demuestran ambas igualdades.

Si $X \sim N(\mu, \sigma) \quad \Rightarrow \quad \frac{(X-\mu)}{\sigma} \sim N(0, 1) \quad \Rightarrow \quad \frac{(X-\mu)^2}{\sigma^2} \sim \chi^2_1$

y en consecuencia :

$$\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$$ por propiedad reproductiva de la distribución ji cuadrado

Se demuestra que:

$$\sum_{i=1}^{n} \frac{(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{o} \quad \text{en forma equivalente: } \frac{(\mathbf{n}-1)\,\mathbf{S}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Además, para una variable $\chi^2_n$ sus parámetros son $\Rightarrow$ $E(\chi^2_n) = n$ y $V(\chi^2_n) = 2n$

y en consecuencia para $\chi^2_{n-1}$ $\Rightarrow$ $E(\chi^2_{n-1}) = n-1$ y $V(\chi^2_{n-1}) = 2(n-1)$ (1)

Reemplazando en (1)

$$E\left[\frac{(n-1)\,S^2}{\sigma^2}\right] = (n-1) \quad \Rightarrow \quad \boxed{E(S^2) = \sigma^2}$$

$$V\left[\frac{(n-1)\,S^2}{\sigma^2}\right] = \frac{(n-1)^2\,V(S^2)}{\sigma^4} = 2(n-1) \quad \Rightarrow \quad \boxed{V(S^2) = \frac{2\sigma^4}{(n-1)}}$$

## 4.- LEY DE LOS GRANDES NUMEROS

En el capítulo 3 se dijo que después de un gran número de repeticiones de una experiencia, la proporción de veces que ocurre un suceso en las n repeticiones (frecuencia relativa), se acerca a la probabilidad de ese suceso. Esto se conoce como "ley de los grandes números" y se puede demostrar matemáticamente a partir de las leyes de probabilidad [8].

- Consideremos una experiencia aleatoria y un suceso A asociado a la misma. Se realizan n repeticiones independientes de la experiencia y se definen las variables:

  $n_A$ : número de veces que ocurre A en las n repeticiones de la experiencia

  $f_A$ : proporción de veces que ocurre A en las n repeticiones de la experiencia

  siendo $f_A = \dfrac{n_A}{n}$

Se conoce que P(A) = p se mantiene constante en las n repeticiones de la experiencia.

Se demuestra que para un número positivo $\varepsilon$:

$$P(\,|f_A - p| < \varepsilon\,) > 1 - \left[\frac{p\,(1-p)}{n\,\varepsilon^2}\right] \qquad \text{y} \qquad \lim_{n\to\infty} P(\,|f_A - p| < \varepsilon\,) = 1$$

es decir, que cuando n tiende a infinito la frecuencia relativa tiende a la probabilidad ( definición frecuencial de probabilidad ).

<u>Demostración</u> :

$n_A \sim Bi(n, p) \qquad E(n_A) = n\,p \qquad \sigma^2(n_A) = n\,p\,(1-p)$

y en consecuencia $\quad E(f_A) = p \qquad \sigma^2(f_A) = \dfrac{p\,(1-p)}{n}$

Aplicando la desigualdad de Tchebychev :

$$P\left[\,|f_A - p| < k\sqrt{\frac{p\,(1-p)}{n}}\,\right] > 1 - \frac{1}{k^2} \qquad (1)$$

Sea $\varepsilon = k\sqrt{\dfrac{p\,(1-p)}{n}}$ $\qquad$ entonces $\quad k^2 = \dfrac{n\,\varepsilon^2}{p\,(1-p)}$

Reemplazando en (1), se obtiene lo que se quería demostrar:

$$P(\,|f_A - p| < \varepsilon\,) > 1 - \left[\frac{p\,(1-p)}{n\,\varepsilon^2}\right]$$

[8] La ley de los grandes números se puede demostrar también a partir del comportamiento del promedio muestral. A medida que el número de observaciones obtenidas aumenta, la media de los valores observados se acerca más y más a $\mu$.